Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation

梁镕深

摘要

使用 deeplabv3+ 实现一个语义分割系统，deeplabv3+ 使用了 aspp，通过多个并行的不同膨胀率的空洞卷积来捕获特征，空洞卷积在不损失信息的前提下，加大了感受野，获取更远距离的特征信息。在 deeplabv3 中未使用解码器，会损失边缘信息，在 deeplabv3+ 中使用了一个 decoder 来获取边缘信息，decoder 结构简单，计算量小，能够逐步获取边缘信息。主干网络使用了 xception 来进行特征提取，加深了网络结构，网络使用了深度可分离卷积，将标准卷积分解为逐通道卷积和逐点卷积，减少了参数量和计算量。

关键词：深度学习；语义分割；encoder-decoder；金字塔结构；深度可分离网络

1 引言

语义分割的目标是为每个图像中的像素分配语义标签，是计算机视觉的基本重要研究方向之一，全卷积神经网络和深度卷积神经网络在基准任务上依赖手工制作的数据集以及标注。在下面复现的工作中，文章使用了空间金字塔和编码器和解码器两种神经网络元素进行语义分割，前者通过不同分辨率下的池化来捕获丰富的上下文信息，而后者能获得清晰的对象边界。

为了捕捉不同尺度的上下文信息，应用了几个不同膨胀率的平行的空洞卷积，这种被称为 aspp（Atrous Spatial Pyramid Pooling）。尽管在最后的特征图里丰富了语义信息，但由于主干网络中池化和步长卷积，语义分割对象的详细边界信息丢失了，本文选择了一个较为有效的编码器和解码器结构来解决这一问题，逐步恢复解码器中的清晰边界，并依旧能保持较快的计算。

最后在 voc 数据集上经过训练达到了 79% 左右的准确率

2 相关工作

全卷积神经网络在语义分割的任务体现了很好的性能，有许多优秀的模型用于语义分割，包括多尺度输入或是采用概率图模型细化分割结果，都取得了较好的效果。

xception 主干网络：能够在减少计算量的同时，保留多尺度的特征，加深层数来获取更高维度的特征提取

2.1 金字塔结构

金字塔结构 Spp 结构在原本的 deeplab[1]或是 pspnet[2]。都已经使用，使用多个平行的不同尺度卷积网络，去捕获不同尺度下的特征作为网络输入，或是使用 aspp 在减小计算量的同时，依旧能捕获不同尺度的特征。

2.2 编码解码结构

Encoder-decoder 编码解码结构，在计算机视觉方面，像是人体模式识别，目标识别，或是语义分割都有较为优秀的应用。编码解码结构有一个编码器来逐步的浓缩特征图去捕捉高纬度下的语义信息。本文在 deeplabv3[3]本身 encoder 部分的基础下，加入一个 decoder 来获取更清晰的语义分割结果。

2.3 深度可分离卷积

深度可分离卷积在卷积层的基础上做了一些改进，将卷积操作分为两个步骤：首先使用深度卷积，然后再进行点卷积。可以减少计算成本和参数量，同时保持较好的性能，这个操作已经我被许多神经网络结构采用，特别是 xception 模型，在 coco 数据集的测试发现，其语义分割任务的准确性和速度都有所改进。

3 本文方法

3.1 本文方法概述

在 deeplabv3+ 的网络结构中主要使用了 Atrous Convolution 和深度可分离网络，并使用了 deeplabv3 中的编码器，并且改进了 xception 网络作为特征分离网络，提高网络计算性能。deeplabv3+ 网络结构如图 1所示：

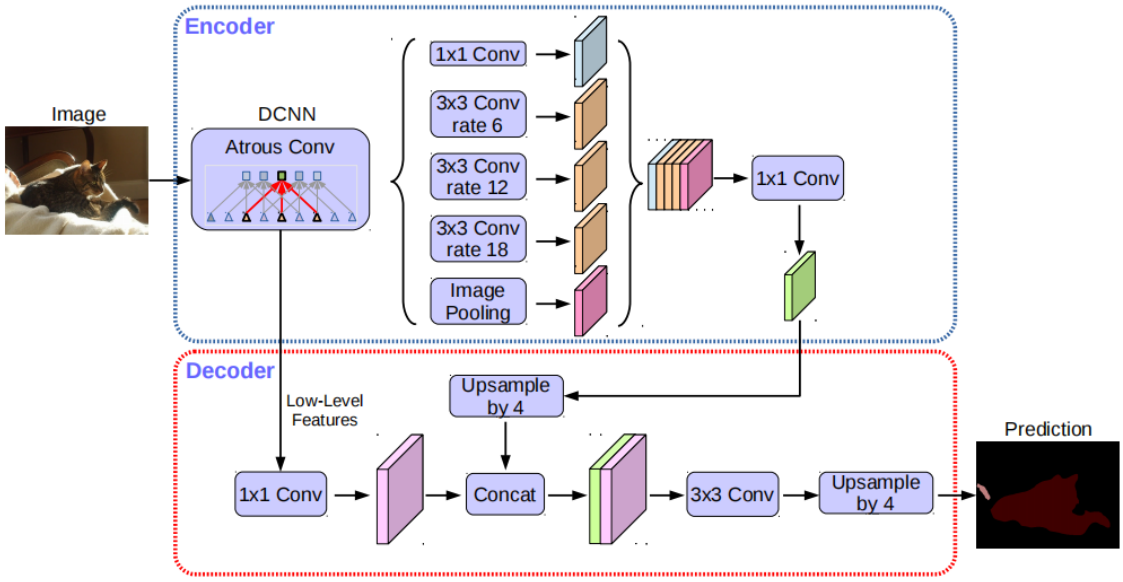

图 1: 网络结构

3.2 编码解码器模块

空洞卷积：空洞卷积可以控制卷积核的感受野大小，在空洞卷积中卷积核的滑步步长不再是固定的，通过设置不同的步长可以捕获不同感受野，捕获不同尺度下的特征。设卷积核为 k，图像为 x，输出为 y，则空洞卷积的计算公式如下：

$$y[i,j] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} k[m,n] \cdot x[i + r \cdot m, j + r \cdot n] \tag{1}$$

其中 r 是卷积核的步长，M 和 N 是卷积核的长度和宽度。当步长为 1 的时候，空洞卷积就是标准卷积。

编码器：deeplabv3+ 沿用了 deeplabv3 的编码器，但将主干网络替换为 xception，保持高性能的同时，通过空洞卷积减少了网络中的参数量，具有更高效率。通过 aspp 增加捕获上下文的能力，通过不

同扩张率的空洞卷积，捕获不同尺度下的特征，最后将生成的特征连接在一起，实现编码的功能。

解码器：通过对编码器生成的输出进行上采样，以便在不损失语义特征的情况下增加图像的分辨率。将编码器的输出进行 4 倍上采样，再与低级特征融合，即保留了边缘信息的同时，拥有高级的语义特征。

4 复现细节

4.1 与已有开源代码对比

源码是使用 tensorflow 实现的网络结构，并且使用 xception 作为主干网络通过空洞卷积，捕获不同尺度下的语义特征，本次复现使用了 pytorch 复现网络，并修改主干网络为 moblienetv2。moblienetv2[4]是由 google 提出的一种轻量级的卷积神经网络，在 moblienetv1[5]的基础上升级，其中使用了残差连接、深度可分离卷积和线性瓶颈来实现搞笑的特征提取和降维，并非线性瓶颈来提高网络的非线性能力。减少网络的复杂度，并保证了网络的性能。

4.2 实验环境搭建

在 linux 操作系统上，使用 anaconda 作为环境管理器，并安装好 pytorch 相关环境，并安装 numpy 和 matplotlib。

4.3 界面分析与使用说明

在安装好的环境下运行 predict.py 就能够简单的测试，输入目录下的图像，如图 2，会在目录下的 output 输出识别结果，如图 3。

```
Configurations:
----------------------------------------------------------------------
|                     keys |                                values|
----------------------------------------------------------------------
|               model_path |       model_data/deeplab_mobilenetv2.pth|
|              num_classes |                                      21|
|                 backbone |                               mobilenet|
|              input_shape |                              [512, 512]|
|         downsample_factor |                                      16|
|                 mix_type |                                       0|
|                     cuda |                                    True|
----------------------------------------------------------------------
Input image filename:img\img.jpg
```

图 2: 识别界面

图 3: 识别结果

4.4 创新点

在复现 deeplabv3+ 原论文的同时，更换了主干网络，分别实现了 xception 和 moblienetv2 的特征分离网络，并同时进行训练。将结果进行比对，发现在使用 moblienetv2 的网络，训练速度快，精确度相差不多

5 实验结果分析

在 voc 数据集上分为训练集和测试集，训练语义分割，并测试精准度，将 miou 和 mPA 作为效果标准，最后能达到将近 80% 的 miou

图 4: 实验结果示意

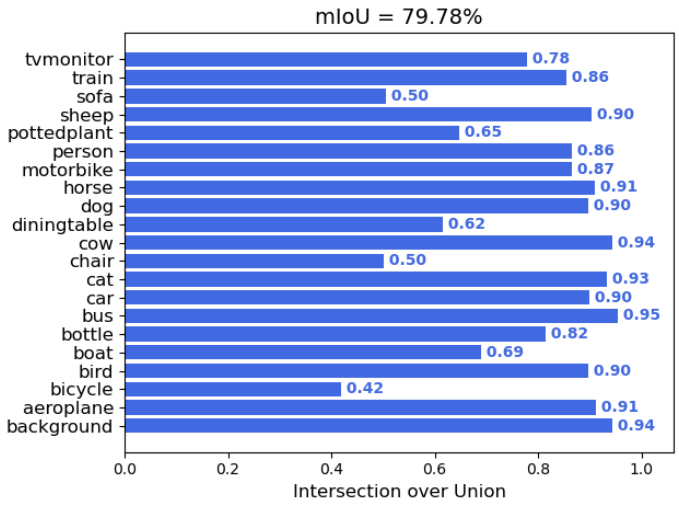

图 5: miou

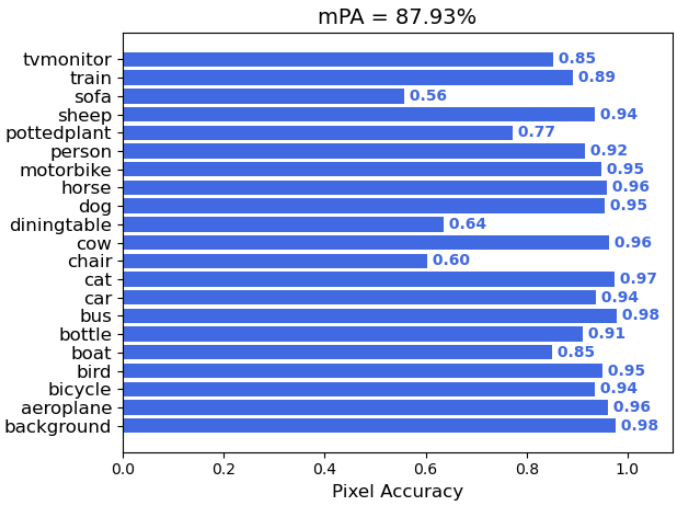

图 6: mPA

6 总结与展望

本文复现了在语义分割领域，有较大进步的 deeplabv3+ 模型，并且在原本 xception 主干网络基础上，额外使用了 moblienetv2 来进行特征捕获，从而达到不同的效果, 语义分割在不同模型都有着不同的进展，本文没能将其他模型的优点吸收来进一步更新 deeplabv3+，达到更好的语义分割效果。今后在不同的语义分割网络模型下能学习不同模型的优点，并融会贯通。

参考文献

[1] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[2] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[3] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.

[4] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[5] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.